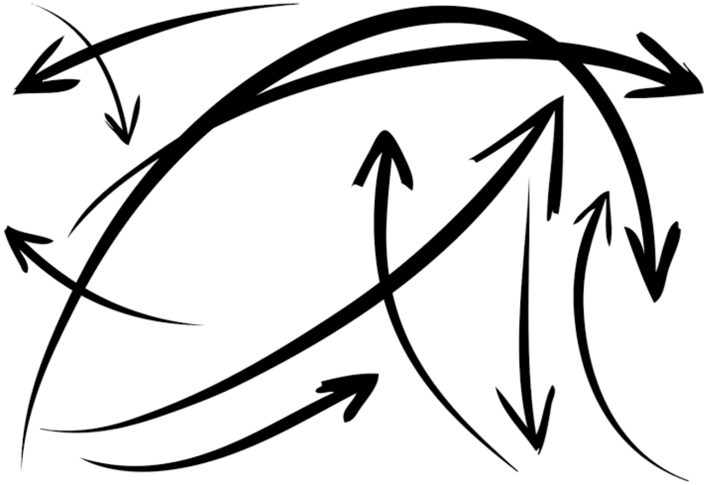
One Vector is not Enough:

Entity-Augmented Distributional Semantics for Discourse Relations

(Ji & Eisenstein 2014)



Mateusz Dubiel

March 31st, 2015

Outline

Background

Discourse relations

Why is recognition of discourse relations difficult?

Methodology

Entity augmented distributional semantics

Prediction of Discourse Relations

Large margin learning framework

Evaluation

Summary

Background information

Discourse relations: bind smaller linguistic units into coherent text.

Discourse recognition types:

- Explicit (He drank some water because he was thirsty)
- Implicit (He drank some water. He was thirsty)

Penn Discourse Treebank: provides large data set of annotations

Automatic identification of implicit discourse relations is very difficult task

- current state-of-art ~40% (Lin et al, 2009)

Reason:

- relations may depend on lower-level elements
- difficult to recover relevant semantics from surface level features

Example 1.

Bob gave Tina the burger.

She was hungry.

Example 1.

Bob gave Tina the burger.

She was hungry.

Bob gave Tina the burger.

<u>Implicit= BECAUSE</u> ***She was hungry.***

Vector Based Representations

Problem: Little surface information to signal the relationship between *burger* and *hungry*

Solution: Discriminatively-trained model, predicting discourse relations as a bilinear combination of vector representations.

- Prediction Matrix and compositional operator are trained to ensure that learned compositional operation produces semantic representations that are useful for discourse

- Although results are positive, purely vector-based approach proves to be not enough

Example 2.

Bob gave Tina the burger.

He was hungry.

Example 2.

Bob gave Tina the burger.

He was hungry.

Bob gave Tina the burger.

<u>Implicit = ALTHOUGH</u>. **He was hungry.**

Vector Based Representations

Problem: Despite the radical difference in meaning, the distributional representation of the second sentence is almost unchanged.

- single vector can’t capture the ways that discourse relations are signaled by entities and their roles.

Vector Based Representations

Problem: Despite the radical difference in meaning, the distributional representation of the second sentence is almost unchanged.

- single vector can't capture the ways that discourse relations are signaled by entities and their roles.

'You can't cram the meaning of whole %&!$# sentence into a single $&!#* vector!' (Mooney 2014)

Solution: compute vector representations for coreferent entity mentions.

Methodology

Entity augmented distributional semantics

The method involves two passes through syntactic structure:

- Upward pass - argument semantics
- Downward pass - entity semantics

Recursive Neural Networks (RNN) - ensure linear relationship between time complexity of the algorithm and input length

Prediction of Discourse Relations

- Relation Identification Model: combines the representations for coreferent mentions

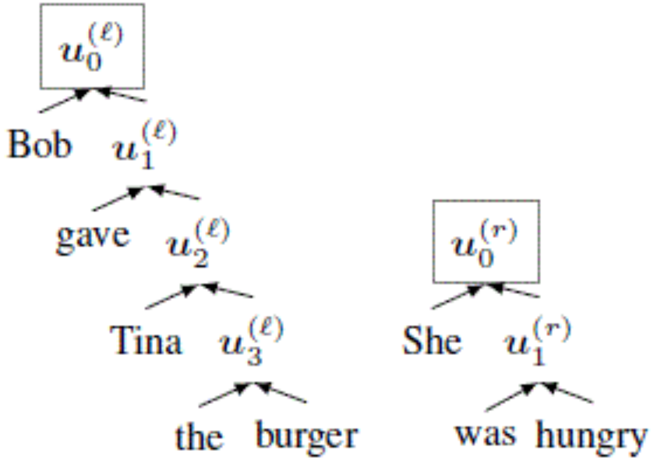
Relation Identification Model

DISCO2 - **dis**tributional, **co**mpositional approach to **disco**urse semantics

Feed-forward compositional model

Upward Pass

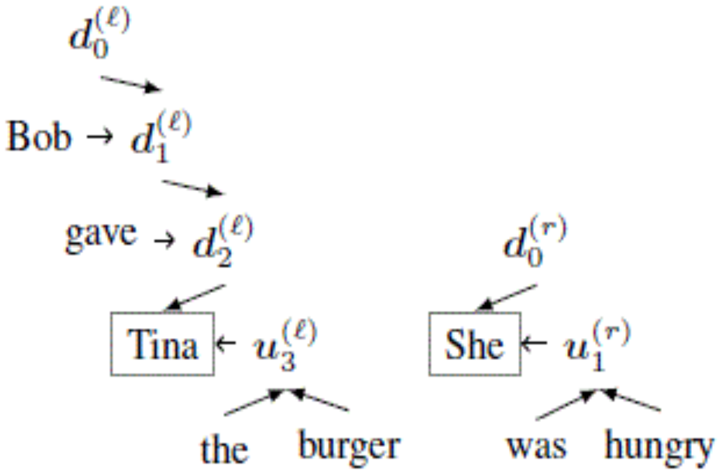
- Finds vector representations for the discourse arguments.
- Each non-terminal in binarised syntactic tree has a 'K - dimensional representation'.
- Computation starts from distributional representation of node's children, bottoming out in individual words.



$$u_i = \tanh\left(\mathrm{U}[u_{l(\mathrm{i})};\mathrm{u}_{r(i)}]\right)$$

Downward Pass

- Adds distributional vectors representing role of co-referent entities.
- The role of constituent *i* is calculated by combining information from neighbouring nodes in parse tree.
- The pass is made by computing the downward vector d_i from the downward vector of the parent $d_{p(i)}$ and the upward vector of the sibling.



$$d_i = \text{tahn}\,(V[d_{p(\mathrm{i})};u_{s(i)}])$$

Relation Identification Model

$$\psi(y) = (u_0^{(m)})^T A_y u_0^{(n)} + \sum_{i,j \epsilon A(m,n)} (d_i^{(m)})^T B_y d_j^{(n)}$$

Decision function: predicts discourse relations between argument pair *(m,n).*

- **Sum of bilinear products**
- **$\hat{y}$ = argmax$_{y \in \mathcal{Y}}$ $\psi(y)$** : gives the predicted relation
- **A_y B_y ($\in \mathbb{R}^{K \times K}$) :** classification parameters
- **b_y :** scalar is used as the bias for term relation *y*
- **A(*m,n*):** set of co-referent entity mentions shared by sentence pair (*m,n*)

Relation Identification Model - Cont.

$$\psi(y) = (u_0^{(m)})^T A_y u_0^{(n)} + \sum_{i,j \epsilon A(m,n)} (d_i^{(m)})^T B_y d_j^{(n)}$$

The decision value *ψ(y)* on relation y is based on :

- Upward discourse vectors at the root $\mathbf{u_0^{(m)}}$ and $\mathbf{u_0^{(n)}}$
- Downward vectors for each pair of aligned entity mentions
- For **A(*m,n*) = Ø**, only upward vector at the root is considered

Relation Identification Model - Cont.

The model is extended to include surface features:

$$\beta^{y^T}\phi_{(m,n)+b_y}$$

Additional **vector $\Phi_{(m,n)}$** - surface features extracted from argument pair *(m,n)*

β_y - marks the classification weight on surface features for relation y

The resulting **decision function**:

$$\psi(y) = (u_0^{(m)})^T A_y u_0^{(n)} + \sum_{i,j\epsilon A(m,n)} (d_i^{(m)})^T B_y d_j^{(n)} \boxed{+ \beta^{y^T}\phi_{(m,n)+b_y})}$$

Implementation

Syntactic structure: Stanford parser used to obtain constituent parse trees of each sentence in PDTB, and binarize all resulting parse trees

Coreference: Berkeley conference system used to extract entities from PDTB

Additional Features: classification model is supplemented using additional surface features i.e.

- ‘lexical features’, ‘constituent parse features’, ‘dependency parse features’, ‘contextual features’

Experiments

Evaluation on PDTB focusing on two types of classification:

- multiclass
- binary

Multiclass classification: evaluation involves predicting the correct discourse relation for each argument pair , from 2nd level of PDTB relations, excluding: ‘Condition’, ‘Pragmatic Condition’, ‘Pragmatic Concession’, ‘Pragmatic Contrast’, and ‘Expression’.

- During training, each argument pair annotated with two relation types considered two training instances.
- During testing, correct if either of two types assigned

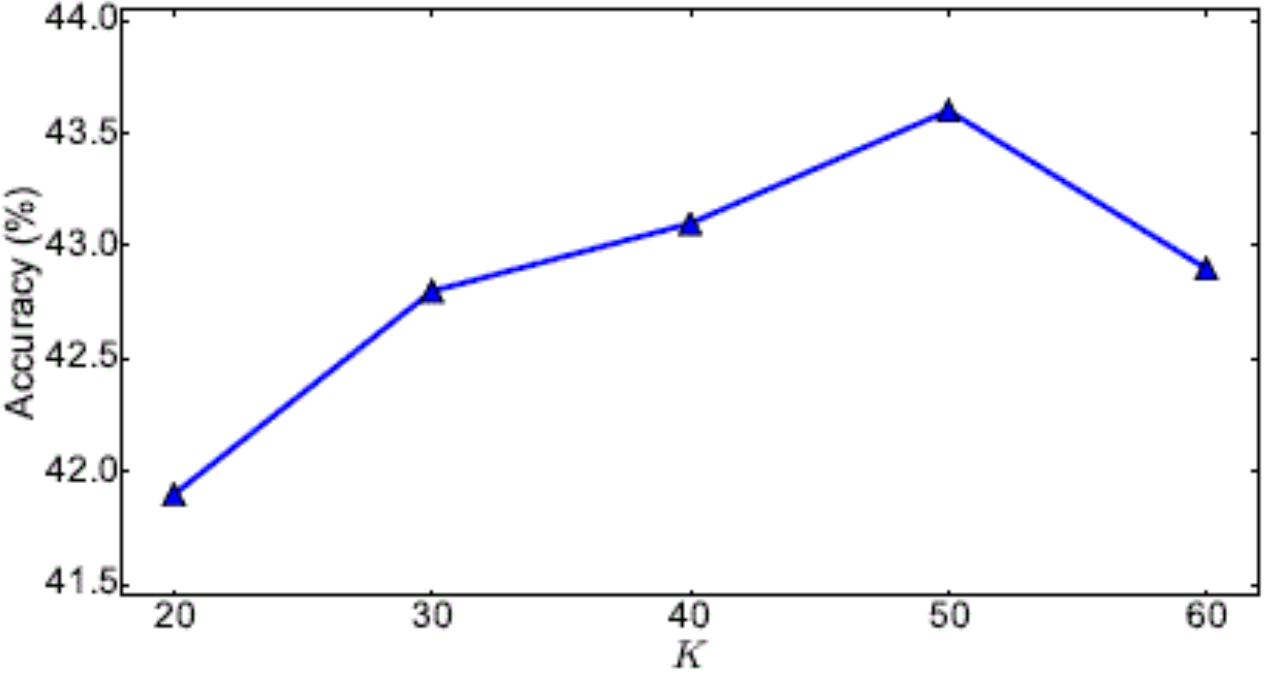
Results - Multiclass identification

Model	+Entity semantics	+Surface features	K	Accuracy(%)
Baseline models				
1. Most common class		No		26.03
2. Additive word representations		No	50	28.73
Prior work				
3. (Lin et al., 2009)		Yes		40.2
Our work				
4. Surface feature model		Yes		39.69
5. DISCO2	No	No	50	36.98
6. DISCO2	Yes	No	50	37.63
7. DISCO2	No	Yes	50	42.53†
8. DISCO2	Yes	Yes	50	**43.56***

* signficantly better than (Lin et al., 2009) with $p < 0.05$

† signficantly better than line 4 with $p < 0.05$

Test set performance - various 'K settings'*



*chosen for distributional representation from a development set

Experiments

Binary classification: evaluation of the four first level relations in PDTB DISCO 2 is applied with downward composition procedure and surface features.

- Four binary classifiers are trained (for each first level discourse relation)
- Sections 2-20 of PDTB (training), 0-1 (development), 21-22 (testing)
- Parameters K,λ, η separately for each classifier by performing a grid search to optimize the F-measure on the developmental data

Results - Binary classification

	Comparison		Contingency		Expansion		Temporal	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Competitive systems								
1. (Pitler et al., 2009)	21.96	56.59	47.13	67.30	76.42	63.62	16.76	63.49
2. (Zhou et al., 2010)	31.79	58.22	47.16	48.96	70.11	54.54	20.30	55.48
3. (Park and Cardie, 2012)	31.32	**74.66**	49.82	72.09	79.22	69.14	26.57	79.32
4. (Biran and McKeown, 2013)	25.40	63.36	46.94	68.09	75.87	62.84	20.23	68.35
Our work								
5. DISCO2	**35.84**	68.45	**51.39**	**74.08**	**79.91**	**69.47**	**26.91**	**86.41**

Evaluation on the first-level discourse-relation identification

Summary

- Predicting discourse relations is fundamentally semantic task.
- Entity-distributional semantics yields significant improvements in implicit relations recognition by including information not only about the semantic arguments but also semantic role of the different entities.
- Recognition of implicit discourse relations still remains one of the unsolved areas of NLP.

Thank you! Any questions/comments?

References:

- *Durrett, G., & Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. In EMNLP (pp. 1971-1982).*
- *Ji, Y., & Eisenstein, J. (2014). One Vector is Not Enough: Entity-Augmented Distributional Semantics for Discourse Relations. arXiv preprint arXiv:1411.6699*.
- *Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 423-430). Association for Computational Linguistics.
- *Lin, Z., Kan, M. Y., & Ng, H. T. (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 343-351)
- Mooney, J, R. (2014). *Semantic parsing: Past, present, and future. Presentation slides from the ACL Workshop on Semantic Parsing*
- *Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 129-136).